Day 3: Linear Regression

Kenneth Benoit and Slava Mikhaylov

Introduction to Data Science and Big Data Analytics

19 August 2015

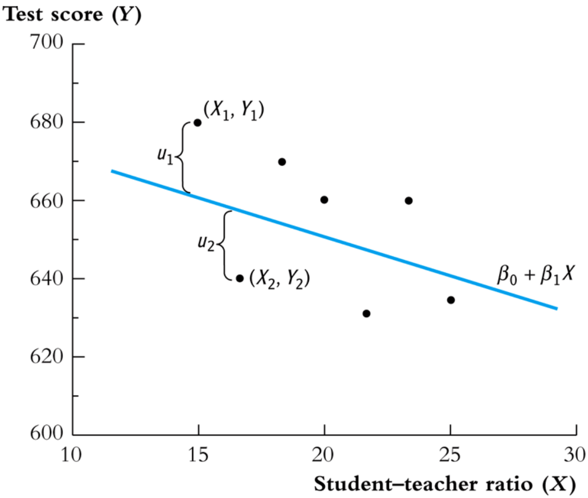
Day 3 Outline

- Regression review
  - Simple linear regression
  - Multiple regression
  - Measures of fit
- Linear regression in R

Regression review

Running example: Class size and student performance

- Consider the problem faced by a school authority:
  - It is considering hiring additional teachers to reduce class; sizes
  - To evaluate this policy the authority would like to know how much student performance will increase as a result of this intervention;
- To help evaluate this policy, you have collected data on test scores and class sizes in 420 school districts in California in 1999.



Test score (Y)
700
680
660
640
620
600
10
15
20
25
30
Student–teacher ratio (X)
(X_1, Y_1)
u_1
u_2
(X_2, Y_2)
$\beta_0 + \beta_1 X$

The linear regression model

- The simplest way to summarize the relationship between two variables is to assume that they are linearly related
- We can express this with the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where:

- y_i is the dependent variable, outcome, or left-hand variable
- x_i is the independent variable, regressor, or right-hand variable
- u_i is the error term
- β_0 and β_1 are parameters to be estimated

The linear regression model (cntd.)

- The Population Regression Function is $\beta_0 + \beta_1$
- The subscript runs over observations $i = 1, \ldots, n$
- In our class size example
  - y_i is the average test score in the school district
  - x_i is the average class size in the school district
  - u_i contains all factors influencing test scores other than class size
  - β_1 is the effect of a one unit change in class size on test scores
  - What does β_0 represent?

Estimating the coefficients of the linear regression model

- If the parameter β_1 were known, it would be very easy to predict the effect of changes in class size.
- How can we estimate the size of β_1 from our data from school districts in California?
- The most widely used approach to estimating the parameters of the linear regression model is the ordinary least squares (OLS) method.

Ordinary Least Squares

- The OLS estimator chooses the regression coefficients so that the estimated regression line is "as close as possible" to the data.
- In particular it minimizes the sum of the squared deviations of the data from the regression line.
- Formally, from all possible β_0 and β_1, it chooses the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following expression:

$$\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{2}$$

- The predicted value of y_i, denoted $\hat{y}_i$, is equal to $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Ordinary least squares (cntd.)

With some algebra one can show that the solution to this minimization problem is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

Why use OLS?

- The OLS estimator is the most popular estimator in applications
- The reason for this is that it has desirable statistical properties under certain assumptions:
  - It is unbiased and consistent
  - Under some additional assumptions it is also the most efficient estimator.
- We examine these conditions next.

Assumptions of the OLS estimator

For the OLS estimator of the parameters β_0 and β_1 to be appropriate three key assumptions have to be satisfied:

1. Conditional (Mean) Independence Assumption: $E(u_i|X_i) = 0$
2. (X_i, Y_i) are i.i.d.: $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d.
3. Large outliers are unlikely

Assumptions of the OLS estimator (cntd.)

- Among these three assumptions Conditional (Mean) Independence is the most critical. Particularly if we want to use the language of causality.
- There are various approaches to deal with violations of i.i.d. (e.g. in time series analysis)
- Assumption 3 can be assessed from the data, but violation can lead to misleading estimation results.

The sampling distribution of the OLS estimator

- The OLS estimator is consistent under conditions listed above.
- We now turn to the efficiency property of the distribution of the OLS estimator.
- It is possible to show that in the absence of heteroskedasticity the variance of the OLS estimator of β_1 is

$$var(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_x^2} \tag{5}$$

The sampling distribution of the OLS estimator (cntd.)

- What is the intuition behind this formula:
  - The larger the variance of the error term the less precise is the estimator of β_1
  - The estimator is more precise as the number of observations n increases
  - Finally, a larger variance of x (for a given σ_u^2) increases the precision of the estimator

The sampling distribution of the OLS estimator (cntd.)

- Even if we know that the OLS estimator of β_1 is consistent and also what its variance is, we still do not know its full distribution.
- It is possible to show that the estimator has a normal distribution if the error term is normally distributed.
- However, fortunately a version of the Central Limit Theorem implies that the estimator will be approximately normally distributed in large samples even if the error term is not normally distributed.
- In practice therefore we are unlikely to rely on the assumption that the error term has a normal distribution to justify that the OLS estimator follows the normal distribution.

Omitted Variable Bias

- So far we have explained the variation in test scores only with the student teacher ratio.
- All other determinants of test scores are therefore included in the error term u.
- Which other determinants of test scores in the school districts of California can you think of?
- Is it problematic to leave these factors included in the error term u?

Omitted Variable Bias (cntd.)

- Omitting a variable from a regression will result in omitted variable bias if two conditions are met:
  - The omitted variable is correlated with the explanatory variable x
  - The omitted variable is a determinant of the dependent variable y
- Which variables may or may not meet these conditions in the test scores example?
- How would we solve this problem?

Multiple regression

- A fairly obvious response to the problem of omitted variable bias is to include further explanatory variables in our regression model (1).
- We could, for example, use an alternative model with two explanatory variables x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{6}$$

- Note that we have now for simplicity dropped the subscript i.
- Is this a general solution to the problem of omitted variable bias?

Multiple regression (cntd.)

To use the OLS estimator to estimate β_1 and β_2 we need one additional assumption:

- There is no perfect multicollinearity between the explanatory variables.

Four Assumptions of the OLS estimator

1. Conditional (Mean) Independence Assumption: $E(u_i|X_i) = 0$
2. (X_i, Y_i) are i.i.d.: $(X_i, Y_i), i = 1, \ldots, n$ are i.i.d.
3. Large outliers are unlikely
4. There is no perfect multicollinearity between the explanatory variables.

Measures of Fit

- How does our model perform? Are we any better than a random guess?
- What proportion of the variation in the dependent variable can be explained by the explanatory variables?

R-squared

- The derivation of the R^2 starts from the identity

$$y_i = \hat{y}_i + \hat{u}_i \tag{7}$$

where

  - y_i is the actual value of the dependent variable for observation i
  - $\hat{y}_i$ is the value of the dependent variable predicted by the regression for observation i
  - $\hat{u}_i$ is the residual and is defined as the deviation of observation i from the regression line, i.e. $\hat{u}_i \equiv y_i - \hat{y}_i$

- Note that the residual $\hat{u}_i$ is not at all the same thing as the error term u_i of the regression model in (1) and (6).

R-squared (cntd.)

- It is possible to show that the total variation in the dependent variable can be decomposed into:

$$TSS = SSR + ESS \tag{8}$$

where

  - TSS (Total sum of squares) equals $\sum_{i=1}^{n}(y_i - \bar{y})^2$
  - ESS (Explained sum of squares) equals $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - SSR (Sum squared residuals) equals $\sum_{i=1}^{n}(y_i - \hat{y})^2$ or simply $\sum_{i=1}^{n}(\hat{u}_i)^2$

R-squared (cntd.)

- The R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{9}$$

  and therefore varies between zero and one.
- How useful is the R^2?
  - A large R^2 most certainly does not imply that a regression represents a causal relationship
  - Similarly, a low R^2 does not by itself mean that a regression is hopeless

The Adjusted R-squared

- R-squared increases when you add a new variable, without corresponding increase in the fit of the model.
- This inflation is corrected through the “adjustment” to the number of independent variables in the model.
- Hence the Adjusted R-squared.

Linear regression in R

Start by loading that ‘MASS‘ and ‘ISLR‘ packages that we will be using throughout this exercise

```r
library(MASS)
library(ISLR)
```

Simple linear regression

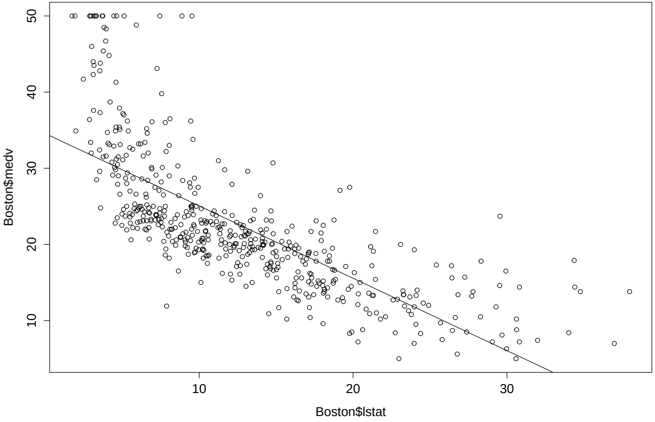
```
lm.fit <- lm(medv ~ lstat, data = Boston)
lm.fit

##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Coefficients:
## (Intercept)        lstat
##       34.55        -0.95

coef(lm.fit)

## (Intercept)       lstat
##  34.5538409  -0.9500494

confint(lm.fit)

##                 2.5 %     97.5 %
## (Intercept) 33.448457 35.6592247
## lstat       -1.026148 -0.8739505
```

Prediction

```
predict(lm.fit, data.frame(lstat = (c(5, 10, 15))), interval = "confidence")

##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461

predict(lm.fit, data.frame(lstat = (c(5, 10, 15))), interval = "prediction")

##        fit       lwr      upr
## 1 29.80359 17.565675 42.04151
## 2 25.05335 12.827626 37.27907
## 3 20.30310  8.077742 32.52846
```

Simple regression plots

```
plot(Boston$lstat, Boston$medv)
abline(lm.fit)
```



Multiple regression

```
lm.fit <- lm(medv ~ lstat + age, data = Boston)
summary(lm.fit)

##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.981  -3.978  -1.283   1.968  23.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
## lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
## age          0.03454    0.01223   2.826  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 503 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.5495
## F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Interaction Terms

```
summary(lm(medv ~ lstat * age, data = Boston))

##
## Call:
## lm(formula = medv ~ lstat * age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.806  -4.045  -1.333   2.085  27.552
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.0885359  1.4698355  24.553  < 2e-16 ***
## lstat       -1.3921168  0.1674555  -8.313 8.78e-16 ***
## age         -0.0007209  0.0198792  -0.036   0.9711
## lstat:age    0.0041560  0.0018518   2.244   0.0252 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 502 degrees of freedom
## Multiple R-squared:  0.5557, Adjusted R-squared:  0.5531
## F-statistic: 209.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

Non-linear Transformations of the Predictors

```r
lm.fit2 <- lm(medv ~ lstat + I(lstat^2), data=Boston)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## lstat       -2.332821   0.123803  -18.84   <2e-16 ***
## I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.524 on 503 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6393
## F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

Adding polynomials

```
lm.fit5 <- lm(medv ~ poly(lstat, 5), data=Boston)
summary(lm.fit5)


##
## Call:
## lm(formula = medv ~ poly(lstat, 5), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5433  -3.1039  -0.7052   2.0844  27.1153
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       22.5328     0.2318  97.197  < 2e-16 ***
## poly(lstat, 5)1 -152.4595     5.2148 -29.236  < 2e-16 ***
## poly(lstat, 5)2   64.2272     5.2148  12.316  < 2e-16 ***
## poly(lstat, 5)3  -27.0511     5.2148  -5.187 3.10e-07 ***
## poly(lstat, 5)4   25.4517     5.2148   4.881 1.42e-06 ***
## poly(lstat, 5)5  -19.2524     5.2148  -3.692 0.000247 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.215 on 500 degrees of freedom
## Multiple R-squared:  0.6817, Adjusted R-squared:  0.6785
## F-statistic: 214.2 on 5 and 500 DF,  p-value: < 2.2e-16
```

Qualitative Predictors

```
lm.fit <- lm(Sales ~ . + Income:Advertising + Price:Age, data = Carseats)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ . + Income:Advertising + Price:Age, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9208 -0.7503  0.0177  0.6754  3.3413
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6.5755654  1.0087470   6.519 2.22e-10 ***
## CompPrice            0.0929371  0.0041183  22.567  < 2e-16 ***
## Income               0.0108940  0.0026044   4.183 3.57e-05 ***
## Advertising          0.0702462  0.0226091   3.107 0.002030 **
## Population           0.0001592  0.0003679   0.433 0.665330
## Price               -0.1008064  0.0074399 -13.549  < 2e-16 ***
## ShelveLocGood        4.8486762  0.1528378  31.724  < 2e-16 ***
## ShelveLocMedium      1.9532620  0.1257682  15.531  < 2e-16 ***
## Age                 -0.0579466  0.0159506  -3.633 0.000318 ***
## Education           -0.0208525  0.0196131  -1.063 0.288361
## UrbanYes             0.1401597  0.1124019   1.247 0.213171
## USYes               -0.1575571  0.1489234  -1.058 0.290729
```

Qualitative Predictors

To examine the coding for the qualitative variables, we can use the "contrasts()" function.

```
contrasts(Carseats$ShelveLoc)

##        Good Medium
## Bad       0      0
## Good      1      0
## Medium    0      1
```

Writing functions

We can define our own functions to wrap a set of 'R' commands in a single call.

```
LoadLibraries <- function() {
    library(ISLR)
    library(MASS)
}
```

And it can then be called like any other R function:

```
LoadLibraries()
```